# Leveraging Concepts in Open Access Publications

Andrea Bertino, Luca Foppiano, Laurent Romary, Pierre Mounier

# Leveraging Concepts in Open Access Publications

**Andrea Bertino[1*], Luca Foppiano[2], Laurent Romary[3], Pierre Mounier[4]**

1 Göttingen State and University Library, Germany

2 ALMAnaCH, Inria, France

3 ALMAnaCH, Inria, France

4 OpenEdition, EHESS. France

*Corresponding Author: bertino@sub.uni-goettingen.de

**Abstract**

This paper addresses the integration of a Named Entity Recognition and Disambiguation (NERD) service within a group of open access (OA) publishing digital platforms and considers its potential impact on both research and scholarly publishing. The software powering this service, called entity-fishing, was initially developed by Inria in the context of the EU FP7 project CENDARI and provides automatic entity recognition and disambiguation using the Wikipedia and Wikidata data sets. The application is distributed with an open-source license, and it has been deployed as a web service in DARIAH's infrastructure hosted by the French Huma-Num. In the paper, we focus on the specific issues related to its integration on five OA platforms specialized in the publication of scholarly monographs in the social sciences and humanities (SSH), as part of the work carried out within the EU H2020 project HIRMEOS (High Integration of Research Monographs in the European Open Science infrastructure). In the first section, we give a brief overview of the current status and evolution of OA publications, considering specifically the challenges that OA monographs are encountering. In the second part, we show how the HIRMEOS project aims to face these challenges by optimizing five OA digital platforms for the publication of monographs from the SSH and ensuring their interoperability. In sections three and four we give a comprehensive description of the entity-fishing service, focusing on its concrete applications in real use cases together with some further possible ideas on how to exploit the annotations generated. We show that entity-fishing annotations can improve both research and the publishing process. In the last chapter, we briefly present further possible application scenarios that could be made available through infrastructural projects.

**Keywords**: Named Entity Recognition and Disambiguation (NERD), Entity-Fishing, Open Access, Monographs, Digital Publishing Platforms

## I Challenges and Perspectives for OA Digital Monographs[1]

Monographs are the privileged means of communication in the humanities and social sciences; more than essays and other shorter publication formats, they enable scientists to deal with differentiated and complex questions in depth. Since monographs require an intensive examination of the subject of study over long periods of time, they contribute to defining the scientific profile of the researcher and thus gain decisive importance for academic careers. Even if not weighted the same in all countries, monographic publications are sometimes a means to acquiring academic qualifications. In Germany, for instance, you may obtain a doctorate or habilitation after the publication of a monograph. However, differentiated monographs are particularly important for the social sciences and the humanities themselves, in that they often open up new research perspectives, i.e. they can be ascribed greater innovative power than articles:

> The process of constructing and writing a book is often a core way to shape the ideas, structure the argument, and work out the relationship between these and the evidence that has emerged from the research process. At their best, monographs provoke debate, can shift paradigms, and provide a focal point for research. It is not surprising […] that the authors of monographs feel a personal connection with the form and content of the works they publish, nor that monographs play a vital role in the careers of many scholars as key markers of esteem and quality. [Crossick, 2015]

Also as collections of different essays, monographs play a unique role in the development of scientific knowledge, by presenting conferences results and allowing interdisciplinary discussion. open access publishing, which increases discoverability and dissemination of research results, can elevate and expand the reach of such discussions. For example, the report by Emery et al. 2017, which has already been broadly discussed online,[2] points out that SpringerLink's OA monographs were downloaded 7 times more frequently on average, received 50% more citations in the following years and were mentioned online 10 times more frequently than their non-open access counterparts.

But despite the increasing recognition of digital books, the dissemination of scholarly monographs in open access is still less common than that of scientific articles. The fact that Directory of Open Access Books (DOAB) was introduced ten years later than the Directory of Open Access Journals (DOAJ) could be considered a reflection of this. DOAB was launched as early as 2012 and, as of November 2018, it contains only about 13,390 books from about 285 publishers -- obviously not broad coverage, given that the American university presses alone publish about 3,000 monographs per year, and BRILL, a heavyweight of the academic book market, around 2,000.

[1] We would like to thank Javier Arias for his valuable comments on this paper.

[2] [Gatti, 2017] shows, among other critical points, a weakness in aggregation of the data concerning the usage of books, for example when downloads of full books and download of chapters are not clearly distinguished.

The publication of monographs in open access is made more difficult by the fact that a complete digitization of the monograph is far more challenging than that of scientific journals and articles, whether in terms of reading habits, reputation gains or storage concepts. In particular, prestige acquisition represents a major challenge in book-oriented disciplines. According to Martin Paul Eve, "Writing a monograph is a substantial commitment of a magnitude many times greater than that of producing a journal article. For this reason, scholars expect a commensurate return on their investment, largely in the form of reputational capital." [Eve, 2014, 120]

Although scientists generally have a positive attitude towards open access publishing, they tend to rely on established legacy publishers when they publish their work, to the detriment of their readership. While the quality evaluation for research results from mathematics, computer science, natural sciences and technology (STEM) is based on apparently objective and not uncontroversial evaluation mechanisms such as journal rankings and impact factors, in SSH much value is still placed on the reputation of the largely highly-specialized publishing houses, to which supposedly better certification of scientific quality is attributed. In order to ensure to the authors gain in reputation through their publications, a certain level of quality certification is required. It is therefore necessary to develop a system for certifying the scientific quality of a publication that can be independent of the prestige of the publisher. For this purpose, it is also essential to enable the collection of bibliometric data that can depict the use and resonance of a particular book. Finally, a more intensive interaction with the publications - whether in the form of open annotations or content mining services - is an added value for authors and should motivate them to choose the open access format for the publication of their monographs. The EU Project HIRMEOS is dedicated to the development of such services which giving new value to Open Access digital monographs.

## II The H2020 HIRMEOS Project

### 2.1 HIRMEOS Aims

The main objective of HIRMEOS (High Integration of Research Monographs in the European Open Science infrastructure) is to enhance five European digital platforms for the publication of open access monographs and to ensure their interoperability. In order to simplify the integration of monographs in the universe of Open Science, several infrastructures and services have been developed in recent years. Consider the following examples: OAPEN (Open Access Publishing in European Networks) (2010); OpenEdition Books (2012); Ubiquity Press (2012); The Directory of Open Access Books (DOAB, 2012); Knowledge Unlatched (2012) and the JSTOR Open Access Book Programme (2016). Nevertheless, the landscape of academic publishing has so far remained highly fragmented among various national, linguistic, and subject-specific contexts. While the publication system in the STEM disciplines is highly standardized and dominated by a few very large players, in the SSH we have—in

addition to some big publishers such as e.g. SpringerNature, Oxford, Cambridge or Chicago University Press— numerous smaller university presses and various online publishing platforms that aggregate titles from several publishers from different national, linguistic and scholarly communities; namely projects like MUSE, The OAPEN Library, Open Edition, JSTOR or ScholarLed. A centralized publishing system, a kind of mega-publishing platform able to overcome the weaknesses of standardization through sensible economies of scale, is still a utopia and would probably be questionable because such an organization would reduce the complexity of the publishing landscape. Nevertheless, it is clear that the current uncoordinated situation represents a strong obstacle to the optimal dissemination of research results in the SSH disciplines and thus has a negative effect on the development of the Open Science model.

Instead of striving for a centralized publishing system, HIRMEOS seeks to achieve horizontal coordination of distributed and already successfully operating platforms. HIRMEOS intends to better integrate open access books, book chapters and anthologies (each peer-reviewed) into the Open Science paradigm and thus enable an explicit step for the humanities and social sciences towards open access. For this step, HIRMEOS deliberately operates as a distributed system in which the homogeneity of the platforms is not achieved simply by using a single software for all of them, but by adopting common standards. The different, independent publishing platforms participating in HIRMEOS are willing to use the same metadata as a result of the project and implement services accordingly. Thus, the publication system remains open to the participation of other platforms, a process which will be simplified by an implementation guide created during the course of the project.

Such an integrated publishing system also strives to support scientific work by simplifying and accelerating basic research activities - the so-called scholarly primitives - functions such as writing, finding, annotating, referencing, assessing, exemplifying, or presenting, but also elementary activities in the digital field such as searching in browsers, connecting digital texts, collecting data, scanning and creating standards of data handling (data practices) (cf. [Unsworth, 2000]).[3] The increasing use of digitized materials in research and learning is also perceived as challenging for our attention and ability to reflect (cf. [Carr, 2008]; [Baron, 2015] and [Pirola, 2017]). The development of new services and tools for digital monographs must therefore always be geared to the concrete needs and practices of researchers and students. Before we describe in detail how HIRMEOS will achieve this, we have to present the structure and method of the project in more detail.

## 2.2 HIRMEOS Partners and Platforms

The HIRMEOS project consortium consists of nine partners in six European countries and one transnational ERIC [ii]:

[3] [Palmer, Teffeau and Pirmann, 2009] speaks of "core scholarly information activities".

1. Centre National de la Recherche Scientifique (CNRS), France

2. Ethniko Idryma Erevnon (NHRF EIE), Greece

3. Stichting OAPEN  Open Access Publishing in European Networks, Netherlands

4. Stiftung Deutsche Geisteswissenschaftliche Institute im Ausland (DGIA), Germany

5. Georg August Universität Göttingen Stiftung Öffentlichen Rechts (UGOE), Germany

6. Ubiquity Press Limited (Ubiquity), United Kingdom

7. Open Book Publishers Community Interest Company (OBP), United Kingdom

8. Digital Research Infrastructure for The Arts and Humanities (DARIAH-EU)

9. Università degli Studi di Torino (UNITO), Italy.

The five digital platforms participating in the project are the following: OpenEdition Books (France), OAPEN (Netherlands), EKT Open Book Press (Greece), Universitätsverlag Göttingen (Germany) and Ubiquity Press (UK). Here is a short description of the platforms:

1. OpenEdition Books (France) is the OpenEdition platform dedicated to open access books.. OpenEdition Books is run by the Center for Open Electronic Publishing (CLEO), the French national infrastructure supported by CNRS, Aix-Marseille University, EHESS and Avignon University. It currently distributes more than 6000 books from 87 publishers. OpenEdition works with Lodel, an open source software developed by the by CNRS-CLEO and disseminates open access books under different models, including the freemium model.

2. The OAPEN Library (Netherlands) is managed by the OAPEN Foundation and, like OpenEdition, aims to provide a highly qualified and certified collection of books. The platform currently presents more than 5000 books from more than 150 publishers. OAPEN also offers publishing houses, libraries, and research funding institutions services in the fields of archiving and long-term archiving, quality certification and dissemination. The OAPEN Library works with XTF, an open source platform developed by the California Digital Library (CDL).

3. EKT Open Book Press (Greece), financed with its own and structural funds, is the service provider for electronic publishing for the Greek National Documentation Center. EKT offers advanced e-Infrastructures and services for institutional partners (universities, research centers, scientific societies, and memory institutions), in particular, to enable the OA publication of peer-reviewed journals, conference proceedings and monographs in the SSH. EKT works with Open Monograph Press (OMP), an Open Source software developed by the Public Knowledge Project to organize the

peer review and editorial processes. OMP can also operate as a website.

4. The Universitätsverlag Göttingen (Germany) is the dedicated publishing house of the Georg-August-Universität Göttingen and is part of the group Electronic Publishing, in which several services and projects of the Niedersächsische und Universitätsbibliothek Göttingen are operated. In addition to other EU projects such as OpenAIRE and FOSTER DFG and BMBF projects on publishing and Open Science, these include advisory services, Open Science Campus activities and various repositories. The university publishing house is managed by an editorial board of the university, which consists of members of all thirteen faculties, ensuring the quality of the publications. The university press publishes about 60 books per year, mainly from the SSH, which are also distributed through print on demand.

5. Ubiquity Press (UK): Ubiquity Press is an open access publisher of peer reviewed journals, academic books and data. Ubiquity provides its own platform and various services. Ubiquity works with RUA, an Open Source application developed by Ubiquity to assist with the monograph publishing life cycle, from submission to both internal and peer review, from copy editing to production and publication.

### 2.3. HIRMEOS Work Package, Data and Services Providers

In order to improve the integration of OA monographs from the SSH, HIRMEOS provides the five publishing platforms with the same standards for various services intended to make the use of open access monographs in the SSH more accessible and attractive to readers. Together with work package (WP) 1, led by the Centre pour l'édition électronique ouverte (CLEO) of CNRS and dedicated to the management and coordination of the project, there are five other work packages dealing with the different technical implementations and a WP taking charge of communication and dissemination of the project results:

WP 2: Identification service. Leader: EKT

WP 3: Named Entities Recognition Service. Leader: DARIAH-EU

WP 4: Certification service. Leader: OAPEN Foundation

WP 5: Annotation service. Leader: Ubiquity Press

WP 6: Metrics service. Leader: Ubiquity Press

WP 7: Community Outreach and Exploitation, Leader: UGOE

In Work Packages 2-6, the use of five different data types is being implemented to improve interoperability between platforms and providers with regard to referencing and indexing services.

These will facilitate the reuse, cross-linking and searchability of content, enable more intensive interaction with users, and strengthen confidence in the quality of published monographs.

### 2.3.1 Metadata for the identification of books and authors

a) All documents published on the platforms are identified by Crossref DOIs. Digital Object Identifier (DOI) technology enables usable, interoperable, and persistent identification of digital objects. DOI technology uses an identification syntax and a network resolution mechanism (Handle System®), a stable and practical infrastructure.

b) If the authors have an ORCID ID, the platforms involved in the project display it next to their name. ORCID (Open Researcher and Contributor ID) is a non-proprietary alphanumeric code for the unique identification of authors. This addresses the problem that the contributions of certain authors can be difficult to recognize since most names of persons are not unique, could change (e.g. in the case of marriages), have cultural differences in the presentation order of names, may contain inconsistent use of first name abbreviations and or utilize different writing systems. The ORCID organization offers an open and independent registry, which is already the de facto standard for the identification of authors of scientific publications.

c) Through FundRef Data, it will be possible to identify the funding institution and the research project behind a specific publication. Publishers can provide financing information for articles and other content using a standard taxonomy of the sponsor's name. A taxonomy of standardized names of the funding agencies is offered by the Open Funder Registry, and associated funding data is then made available via Crossref search interfaces and APIs for sponsors and other interested parties.

All in all, the incorporation of these standards across the platforms should significantly improve the findability of open access monographs, which is not always optimal today (cf. [McCollough, 2017]).

### 2.3.2 Peer-Review Certification

The majority of scientific monographs undergo intensive quality assurance and evaluation procedures, which are, however, less standardized in the SSH than in other disciplines. Regardless of which review procedure would be optimal for monographs, HIRMEOS is developing a certification system that categorizes and standardizes review procedures. In this way, users can immediately recognize which review procedure a publication has undergone. A peer review certificate and an OA license certificate will be added to each document published on the five platforms.

The peer review and OA license certificates are delivered to DOAB's various partners. DOAB offers a quality-controlled list of peer-reviewed open access monographs (including book chapters) and book publishers. By developing this quality-controlled list, DOAB enables researchers, libraries, and discovery services to easily identify and search peer reviewed open access monographs, improving

the discoverability, access, and use of monographs around the world. After an application process, publishers that meet the DOAB peer review and open access requirements and have the corresponding licenses are listed in DOAB and can upload the metadata of their open access books. Such metadata can then be disseminated through the OAI-PMH protocol implemented by third-party providers such as libraries and search services, thus improving the findability of books.

DOAB also offers an automated upload service for OA books from trusted OA platforms. Current platforms using this service are the OAPEN Foundation and SciELO (Scientific Electronic Library Online), the latter being mainly run from Brazil in Latin America. The DOAB certification service includes a classification system and also allows certified publishers and publishing platforms to collect DOAB certification and icons through the DOAB API. Certified publishers and publishing platforms that meet DOAB's requirements agree to the conditions of DOAB certification and commit to passing an audit to verify their peer review procedures.

### 2.3.3 Annotation of digital monographs

HIRMEOS has already made it possible for users of monographs across the five platforms to add annotations using the open annotation tool Hypothes.is.

Hypothes.is is an open platform for discussion on the web that allows annotations to be written at sentence or word level, such as criticism or notes on news, blogs, scientific articles, books, terms of use, campaign initiatives, legislative procedures, and more. Hypothes.is is based on an open source JavaScript library and annotation standards developed by the W3C Web Annotation Working Group. Hypothes.is has established broad partnerships with developers, publishers, academic institutions, researchers, and individuals to provide a platform for next-generation read-write web applications. The Hypothes.is software is developed by a non-profit organization, financed by the generosity of the the Knight, Mellon, Shuttleworth, Sloan and Helmsley Foundations. A coalition of over 60 scientific publishers, including PLOS, Wiley, Oxford University Press, support Hypothes.is through the Annotating All Knowledge Initiative.

### 2.3.4 Metadata for metrics and legacy metrics

The measurement of impact and resonance presents specific challenges for open access monographs. Keeping track of current downloads, readership, and reuse across multiple platforms is difficult, but important if one wants to understand and track their reach. The use of alternative metrics (Altmetrics), which measure the number of mentions of a document in social networks and other kinds of publications, has also increased significantly in recent years because it helps to better understand the impact of scientific publications by documenting the resonance of scientific content in broader communities and beyond the specific academic context.

HIRMEOS partners Ubiquity Press and Open Books Publishers are enabling the platforms to collect

usage metrics and alternative metrics and to display them directly on the documents. The software collecting altmetrics will be hosted by Ubiquity Press and all code, APIs, and standards will be published Open Source. The altmetrics service will record the following measures for books: tweets, Facebook shares, Wikipedia quotes, and annotations. The service is designed to operate on a daily basis; it can therefore also make this data available in chronological order. Since the data sources often change their access methods, licensing and conditions, this is a maintenance-intensive system. Ubiquity Press will host, manage and maintain this system and make it available as a service to other platforms. The entire source code will be available under an Open Source license (MIT) for the participating platforms and the wider community to use, reuse and expand.

Open Book Publishers, on the other hand, has developed its own software capable of collecting, processing, and standardizing usage metrics from third-party platforms. A key component of their software stack is a database and API containing mappings of the different identifiers used by each platform (e.g. URLs, ISBNs, DOIs) to the pertinent work, whether it is a monograph, a chapter, an image, or any digital object constituting a publication on its own. The service then extracts usage data from the platform in question and uses the identifiers database to enforce the use of a particular type of identifier, such a DOI, i.e., then translates the identifiers used by the reporting platform to the one desired by the data collector. The data collected is stored in the form of events, each of them recording the measure collected, the timestamp, the identifier of the work affected by the event, and the number of times the event was repeated, e.g., there were four downloads of this book in this platform. The standardization process not only normalizes identifiers, it also tags each event with a URI identifying the measurement it represents (e.g. views), the platform reporting the event (e.g. Google Books), and provides a location of the definition of the measure for further user-friendly description. The format described is provided as an open standard, allowing its adoption by any platform or publisher wishing to collect and display usage data from any distributing platform. Similarly, the software developed to collect metrics is provided Open Source, and containerized, as to allow its reuse.

## III Entity Resolution Service

### 3.1 Text Mining and entity-recognition: state of art and most relevant tools

With the digital information explosion over the last few decades, the extraction and resolution of entities has been studied extensively (cf. [Milne et al., 2007]) and has become a crucial task in large-scale text and data mining (TDM) activities. Entity extraction and resolution is the task of determining the identity of entities mentioned in a text against a knowledge base representing the reality of the domain under consideration. This could be the recognition of generic Named Entities suitable for general purpose subjects, like people, location, organizations and so on, but also the resolution of

specialist entities in different domains.

Entity-fishing addresses these needs and provides a generic service for entity extraction and disambiguation (NERD) against Wikidata, supporting possible further adaptation for application to specialist domains. This allows it to be independent of a particular framework and usage scenario for maximum reuse.

Entity-fishing offers close to state-of-the-art accuracy (as compared with other NERD systems). The accuracy f-score for disambiguation is currently between 76.5 and 89.1 on standard datasets (ACE2004, AIDA-CONLL-testb, AQUAINT, MSNBC)[1].

| | Priors | entity-fishing | Wikifier | DoSeR | AIDA | Spotlight | Babelfy | WAT | (Ganea & Hofmann, 2017) |
|---|---|---|---|---|---|---|---|---|---|
| **ACE2004** | 83.1 | 83.5 | 83.4 | **90.7** | 81.5 | 71.3 | 56.1 | 80.0 | 88.5 |
| **AIDA-CONLL-testb** | 66.1 | 76.5 | 77.7 | 78.4 | 77.4 | 59.3 | 59.2 | 84.3 | **92.2** |
| **AQUAINT** | 80.3 | **89.1** | 86.2 | 84.2 | 53.2 | 71.3 | 65.2 | 76.8 | 88.5 |
| **MSNBC** | 71.1 | 86.7 | 85.1 | 91.1 | 78.2 | 51.1 | 60.7 | 77.7 | **93.7** |

The objective is to provide a generic service having a steady throughput of 500-1000 words per second or one PDF page of a scientific article in 1-2 seconds on a medium range (4CPU, 3Gb Ram) Linux server.

The entity-fishing API allows the processing of different input (raw or partially annotated texts, PDFs, search queries), different languages and different formats. Entity-fishing employs supervised Machine Learning algorithms for both the recognition and the disambiguation tasks, using training data generated from Wikipedia article structures (cf. [Milne, 2008]).

## 3.2. The entity-fishing service

Deployed as part of the national infrastructure Huma-Num in France, this service provides an efficient state-of-the-art implementation coupled with standardized interfaces, allowing an easy deployment in a variety of potential digital humanities contexts.

Entity-fishing implements entity extraction and disambiguation against Wikipedia and Wikidata entries. The service is accessible through a REST API which allows simple and seamless integration, language independent and stable convention and a widely used service oriented architecture (SOA) design.

The interface implements a variety of functionalities, like language recognition, sentence

segmentation, and modules for accessing and looking up concepts in the knowledge base. The API itself integrates more advanced contextual parameterization or ranked outputs, allowing for the resilient integration in various possible use cases. The representation is also compliant with the Web Annotation Data Model (WADM).

The details of how entity-fishing works are outside the scope of this paper. For more information on this subject, we recommend reading the paper [Foppiano et al., 2018].

## IV Entity-fishing integration: applications for scholarly publishing

In this section we will present the integration of the entity-fishing service within the open access digital publishers' infrastructures. We will focus on the use cases, which were the objects of the implementations.

The work carried out during the project was supervised and measured by different levels of increasing complexity (from the access to the API to the "creation" of new services using the generated data), and the implementation was driven by each partner's needs. This flexibility had two effects, on one side it gave us the opportunity to expand the use case "portfolio" of the service due to the heterogeneity of the platforms, and on the other hand it necessitated understanding of what each service was able to achieve and how the resulting data could be used. This second aspect was critical, mostly due to the lack of skills and expertise in within the humanistic domain in regards to TDM.

Seeing how the service was received and what were the aspects with higher learning curves was an interesting experiment.

This section is divided into two parts, the first part illustrates the use cases specifically implemented in the digital library infrastructure, including links to their production environments; the second part will show some ideas and further improvements.

### 4.1 Use cases

Before discussing all the scenarios in detail, it is informative to detail the amount of data each partner had to work with:

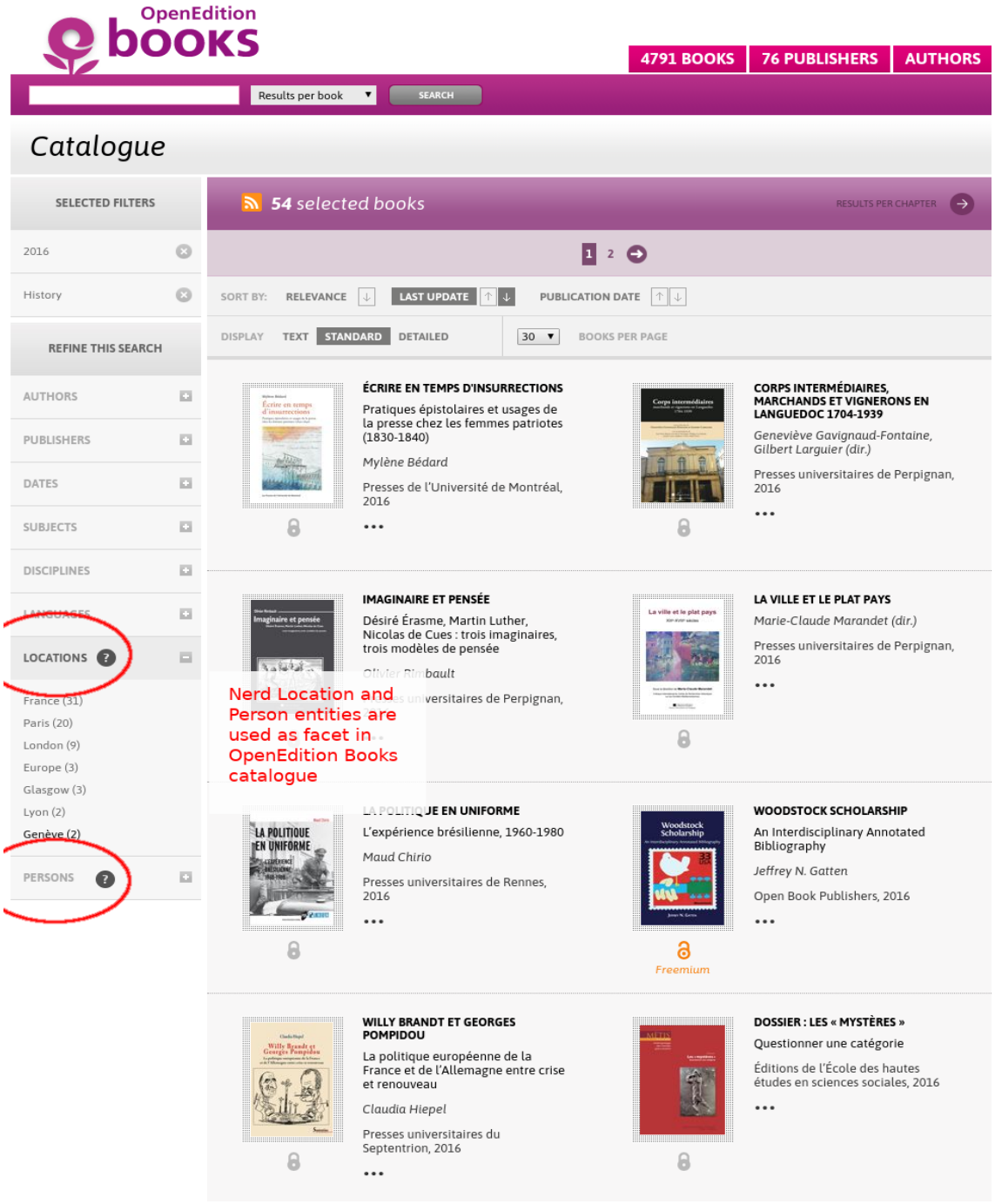
- 4000 books in English and French from Open Edition
- 2000 titles in English and German from OAPEN
- 162 books in English from Ubiquity Press
- 765 books (606 in German, 159 in English) from UGOE
- EKT had just one book in English, the rest was in Greek

Although the preferred publishing format varies from platform to platform, whether it is XML, PDF, or HTML, *entity-fishing* is able to handle XML, PDF, and plain text. However, as each organization uses a different publishing platform, the design and methodology used in the implementation had to be tailored to each case.

One of the most frequently implemented use cases was the improvement of the platform's search interface by using entities extracted from the library content. This was done by extracting specific Named Entities and enabling users to filter their search with these complementary parameters.

OpenEdition extended their Books Catalogue[4] by adding two additional facets to filter books by entities of type PERSON and LOCATION.

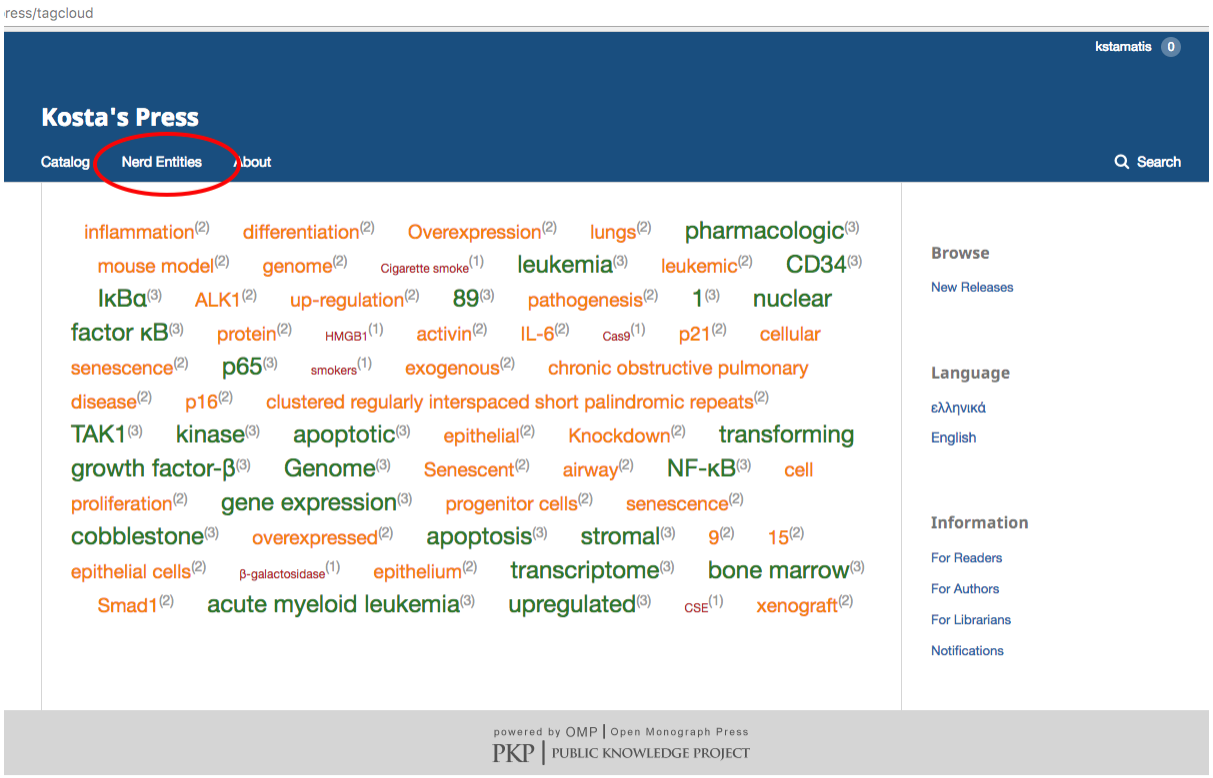
[4] http://books.openedition.org/catalogue




Along the same lines, Goettingen State Library also extracted mentions of organizations from the library corpus in their book catalog[5]:

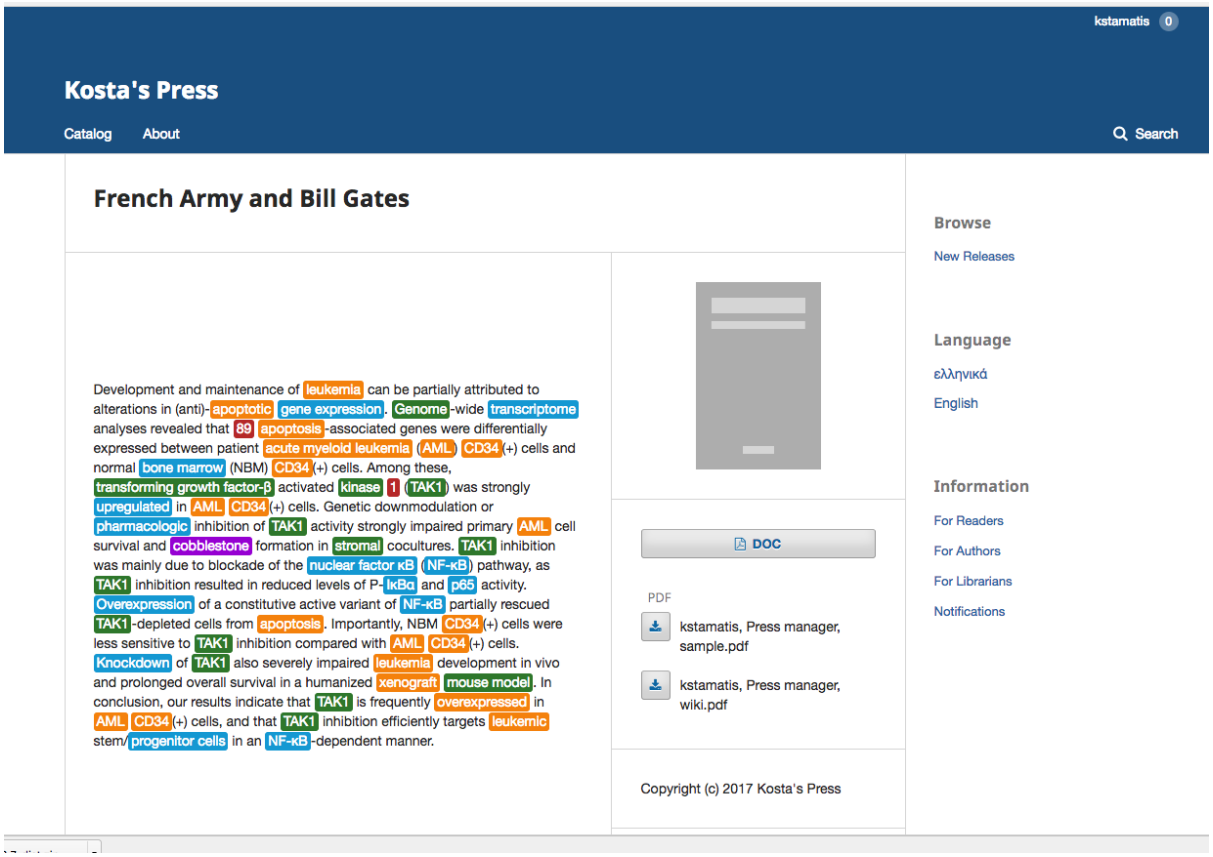
[5] https://www.univerlag.uni-goettingen.de/handle/3/Goettingen_studies_in_cultural_property_series

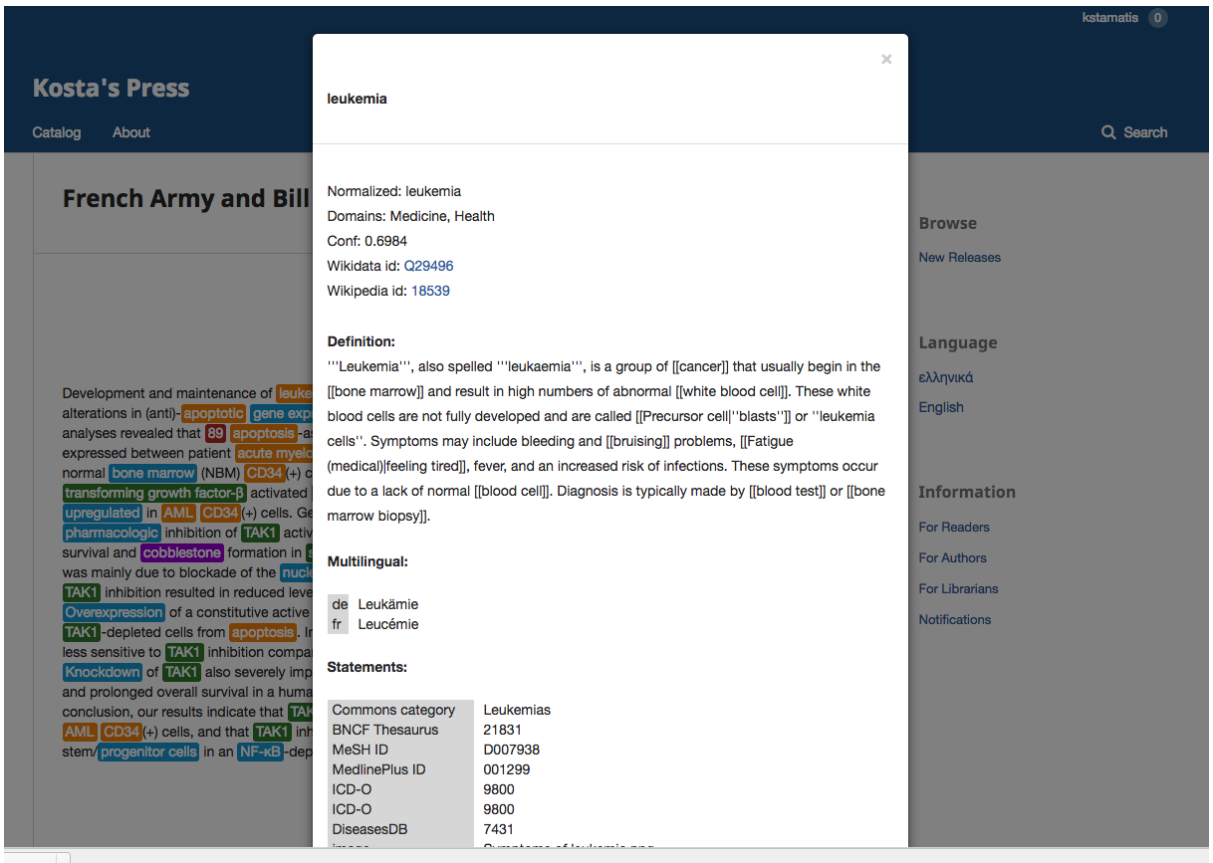

Once the annotation data is collected and stored, it can also be used for automatic generation of word clouds at the repository level, with words displayed according to their importance (relevance, frequency, etc.). Users are then able to access the most relevant concepts at library level. The underlying data is effectively the same used for facet searching, but with a different visualization.




A more interesting evolution, aiming to dramatically improve the search quality, would be the implementation of the word cloud as a facet. We can imagine the cloud evolving at each search or filtering as the user narrows the search space, providing clearer insight into the search results.

A second generic aspect the annotation can enable is the visualization of concepts within the text. The EKT (National Documentation Center) implemented concept visualization for abstracts and titles, allowing the highlighting of the most relevant information and the prospect of learning about it using the Wikipedia definition of the concept directly in the page.


kstamatis 0
Kosta's Press
Catalog
About
Search
French Army and Bill Gates
Development and maintenance of leukemia can be partially attributed to alterations in (anti)-apoptotic gene expression. Genome-wide transcriptome analyses revealed that 89 apoptosis-associated genes were differentially expressed between patient acute myeloid leukemia (AML) CD34(+) cells and normal bone marrow (NBM) CD34(+) cells. Among these, transforming growth factor-β activated kinase 1 (TAK1) was strongly upregulated in AML CD34(+) cells. Genetic downmodulation or pharmacologic inhibition of TAK1 activity strongly impaired primary AML cell survival and cobblestone formation in stromal cocultures. TAK1 inhibition was mainly due to blockade of the nuclear factor κB (NF-κB) pathway, as TAK1 inhibition resulted in reduced levels of P-IκBα and p65 activity. Overexpression of a constitutive active variant of NF-κB partially rescued TAK1-depleted cells from apoptosis. Importantly, NBM CD34(+) cells were less sensitive to TAK1 inhibition compared with AML CD34(+) cells. Knockdown of TAK1 also severely impaired leukemia development in vivo and prolonged overall survival in a humanized xenograft mouse model. In conclusion, our results indicate that TAK1 is frequently overexpressed in AML CD34(+) cells, and that TAK1 inhibition efficiently targets leukemic stem/progenitor cells in an NF-κB-dependent manner.
DOC
PDF
kstamatis, Press manager, sample.pdf
kstamatis, Press manager, wiki.pdf
Copyright (c) 2017 Kosta's Press
Browse
New Releases
Language
ελληνικά
English
Information
For Readers
For Authors
For Librarians
Notifications


kstamatis 0
Kosta's Press
Catalog
About
Search
leukemia
Normalized: leukemia
Domains: Medicine, Health
Conf: 0.6984
Wikidata id: Q29496
Wikipedia id: 18539
Definition:
'''Leukemia''', also spelled '''leukaemia''', is a group of [[cancer]] that usually begin in the [[bone marrow]] and result in high numbers of abnormal [[white blood cell|]]. These white blood cells are not fully developed and are called [[Precursor cell|''blasts'']] or ''leukemia cells''. Symptoms may include bleeding and [[bruising]] problems, [[Fatigue (medical)|feeling tired]], fever, and an increased risk of infections. These symptoms occur due to a lack of normal [[blood cell|]]. Diagnosis is typically made by [[blood test]] or [[bone marrow biopsy]].
Multilingual:
de Leukämie
fr Leucémie
Statements:
Commons category Leukemias
BNCF Thesaurus 21831
MeSH ID D007938
MedlinePlus ID 001299
ICD-O 9800
ICD-O 9800
DiseasesDB 7431
Browse
New Releases
Language
ελληνικά
English
Information
For Readers
For Authors
For Librarians
Notifications

Annotations displayed in this way could help to achieve a better and quicker understanding of the specific and disciplinary language of the search results thus encouraging non-specialists to use them. Exploiting annotations in such a way could potentially foster interdisciplinary research.

## New scenarios and ideas

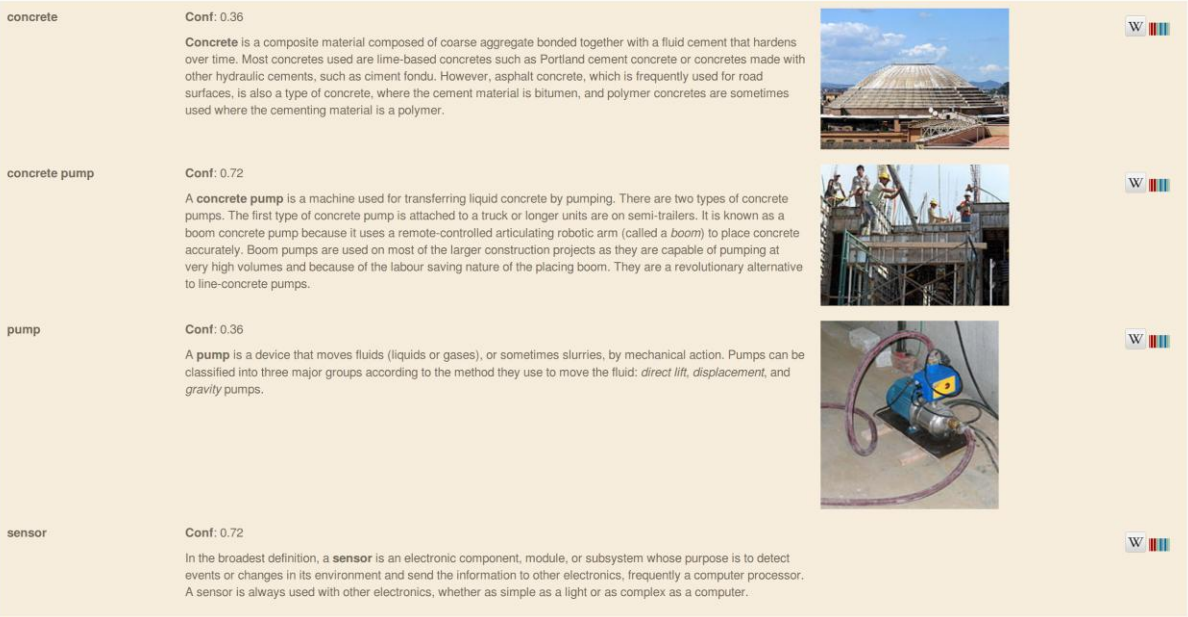
The use cases implemented during the project were the first step towards context-aware tools for extracting information from the full text.

In this section the goal is to propose some improvements of what has already been done and to provide further ideas and interesting applications. The main aspects are search, visualization, and clustering.

### 4.2 Search

Improving search results is one of the biggest areas where annotation of concepts and entities can bring the most benefits. A user can further narrow search results by selecting the different meanings the word may have, for example someone searching for *Washington* could mean the city, the state, or the various persons carrying that name.

*Entity-fishing* already supports a "query disambiguation" mode that allows the user to propose which concept they are interested in. For example, searching *concrete pump sensor* would result in the following disambiguation entries[6]:




The user selection will generate additional options to direct the search to match the correct concept

[6] This screenshot is taken from the official github page of *entity-fishing*: http://github.com/kermitt2/entity-fishing

from among several possibilities.

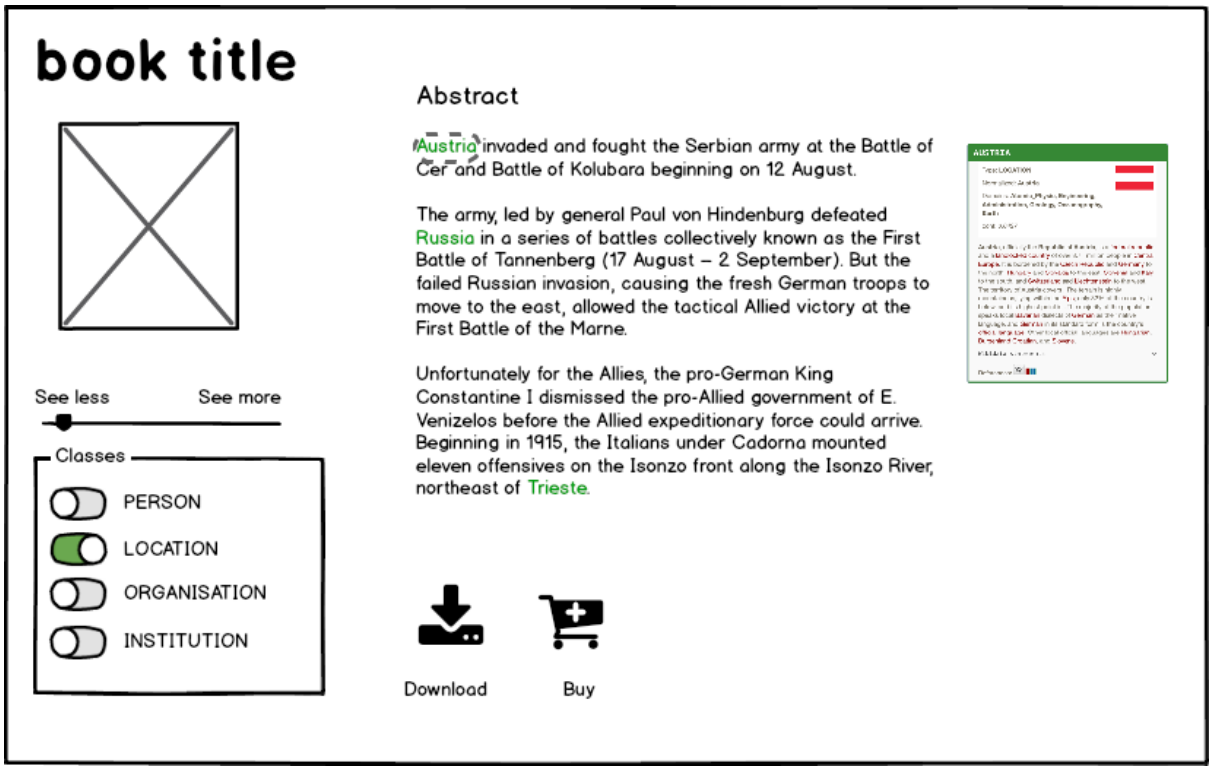
### 4,3 Visualisation

The visualization of entities is a tricky subject because its acceptance by users is not uniform. While some enjoy colorful text, others might find the effect disturbing. It is therefore important to implement this functionality in the right way: letting the user know that the functionality exists but with a relatively light weight integration. For example, the safe approach would be to not display any annotation by default, or to display very few (perhaps the ones with the highest relevance) and to offer the user of the option to discover / highlight more.

With regards to implementation, we offer several options:

- a slider would allow users to decrease or increase the amount of annotations displayed by filtering out entities according to confidence score,
- a set of checkboxes by entity type would allow users to select only entities of certain types, like PEOPLE, but also by domain, for example Biology, Chemistry, Engineering, etc.

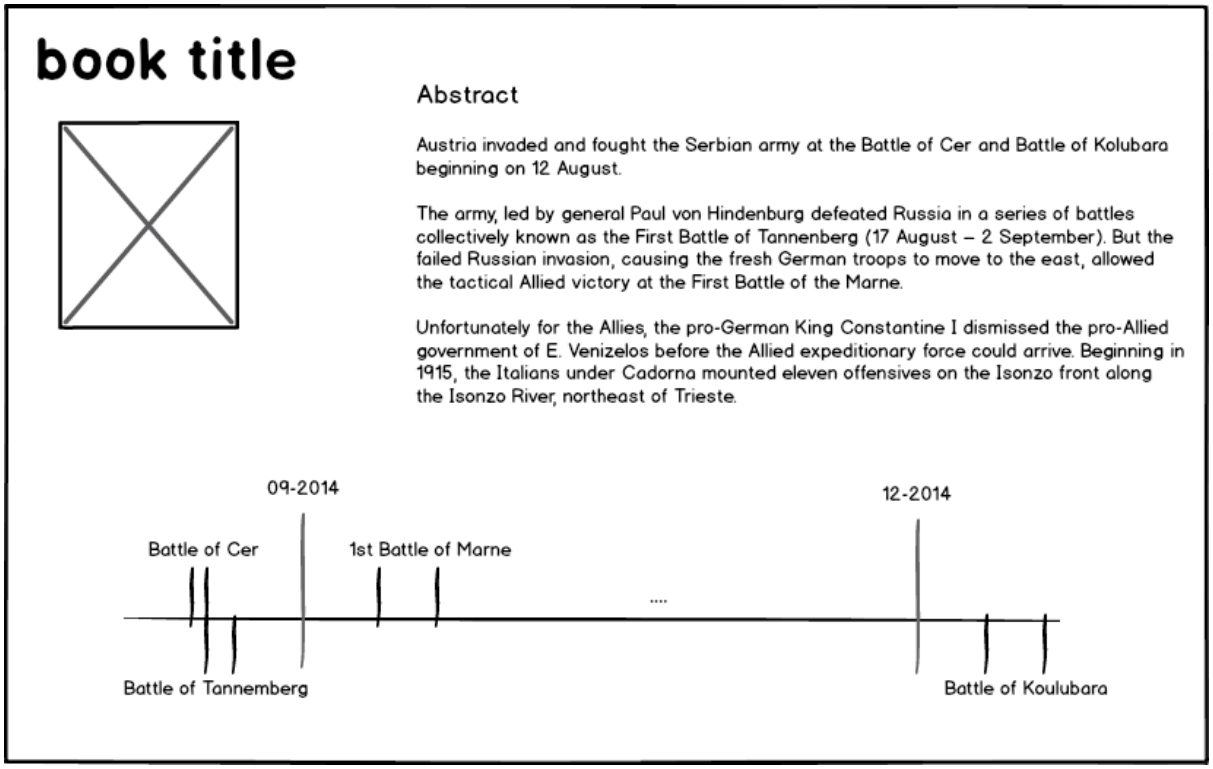
These are illustrated in the mockup below.




As a side note, the confidence score is one of the most important aspects to consider when analyzing the output data. We stress this point particularly when the objective is to have very precise results and the recall is not important. Finding a good balance can be left to post-processing.

Another challenging idea is the possibility to extract all events and temporal expressions in order to build a timeline visualization graph at the collection or book level. At book level it is interesting to have a quick summary of the content of the title, while at repository level it offers interesting possibilities for matching books that discuss for example the same historical period, or to find connections between different documents and books.

This concept is illustrated in the following mockup.




**Clustering**

Clustering allows grouping of books by their content, this enables automatically generated collections with links to books. These prospects could reveal new relationships and boost the discovering of certain books on the basis of their content.

Classic recommendation systems attempt to suggest additional articles or products the user might be interested in by exploiting the user's purchase or navigation history. This could have a big impact on the dissemination of open access monograph catalogs.

A simpler approach (which would not require collecting any user data) would be to process the g annotations with clustering techniques. This would enable more cluster configurations, exploiting different aspects (domains, similarity, frequency, etc.) of the collected annotations.

These ideas and scenarios are applicable not only to other publishing platforms in SSH but potentially to an open access repository.

## V Possible application on a discovery platform

Entity-fishing could also be integrated in the process of feature generation as recommended by CORE (https://core.ac.uk/), extracting concepts to be used as keywords for linking research outputs together.

Indexing a corpus of publications by extracting keywords via entity-fishing can significantly increase the discoverability of these publications. However, indexing only makes sense if it includes a very high number of digital objects. In light of this, OPERAS, the European infrastructure dedicated to open scholarly communication in the humanities and social sciences, identified in its design study (cf. [OPERAS Consortium, 2017]) the need for a discovery platform at the European level that could help European researchers in these disciplines discover content relevant to their research across the widely scattered servers that host content throughout Europe. Based on the French Huma-Num Isidore platform,[7] this future discovery platform will be multilingual and index not only publications (journal articles and monographs), but also other forms of scholarly communication like blog posts and conference programs, and research data as well. One of the key functionalities of this platform will be, of course, its ability to crosslink content alongside different conceptual lines and vocabularies to increase discoverability for the user through suggestions and recommendations. In this context the *entity-fishing* service appears to be one of the central elements of the platform as it will enable the user to navigate across the wide variety of content indexed by the platform through person names, locations, and concepts.

[7] https://www.rechercheisidore.fr/

## References


Baron N. Words Onscreen: *The Fate of Reading in a Digital World*. Oxford University Press (Oxford), 2015.

Carr N. *Is Google Making Us Stupid?* 2015. https://www.theatlantic.com/magazine/archive/2008/07/is-google-making-us-stupid/306868.

Crossick G.: *Monographs and Open Access*. A report to HEFCE, HEFCE 2015. http://www.hefce.ac.uk/pubs/rereports/year/2015/monographs

Cucerzan S. *Large-scale named entity disambiguation based on wikipedia data*. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007: 708-716.

Emery C., Lucraft M., Morka A.; Pyne R. The OA Effect: How does open access affect the usage of scholarly books? 2017. https://www.springernature.com/gp/open-research/journals-books/books/the-oa-effect

Eve MP. Open access publishing and scholarly communications in non-scientific disciplines. *Online Information Review*. 2015; 39(5): 717–732.

Foppiano L., Romary L. *entity-fishing: a DARIAH entity recognition and disambiguation service*. Communication at the conference *Digital Scholarship in the Humanities* , Sep. 2018, Tokyo, Japan. https://hal.inria.fr/hal-01812100

Gatti R. *Handle with Care: pitfalls in analysing book usage data*, 2017. https://rupertgatti.wordpress.com/2017/12/11/handle-with-care-pitfalls-in-analysing-bookusage-data.

Lopez P., Meyer A., Romary L. *CENDARI Virtual Research Environment & Named Entity Recognition techniques.* Poster at Grenzen überschreiten – Digitale Geisteswissenschaft heute und morgen. Feb 2014; Berlin, Germany. https://hal.inria.fr/hal-01577975

Lopez P., *entity-fishing*, Presentation at WiidataConf; 2017: . https://grobid.s3.amazonaws.com/presentations/29-10-2017.pdf

McCollough, Aaron. *"Does It Make a Sound: Are Open Access Monographs Discoverable in Library Catalogs?"* portal: Libraries and the Academy, 2017; 17(1): 179-194.

Milne DN., Witten IH., Nichols DN. *A knowledge-based search engine powered by wikipedia.* Proceedings of the 16th ACM conference on information and knowledge management. ACM, New York, NY, USA, 2007: 445-454.

Milne D., Witten IH.. *Learning to link with wikipedia*. Proceedings of the 17th ACM conference on Information and knowledge management. ACM; New York, NY, USA, 2008: 509–518.

OPERAS Consortium. *Operas design study*, October 2017. https://zenodo.org/record/1009544#.XDxZN1VKhtQ.

Palmer CL., Teffeau LC., Pirmann CM. *Scholarly information practices in the online environment: Themes from*

*the literature and implications for library service development*, Report commissioned by OCLC Research. Published online at: www.oclc.org/programs/publications/reports/2009-02.pdf

Pirola S. *The Academic Book of the Future and its Readers*.UCL Press (London), 2017.

Unsworth J. *Scholarly Primitives: What methods do humanities researchers have in common, and how might our tools reflect this?* London, 2000. http://people.virginia.edu/~jmu2m/Kings.5-00/primitives.html